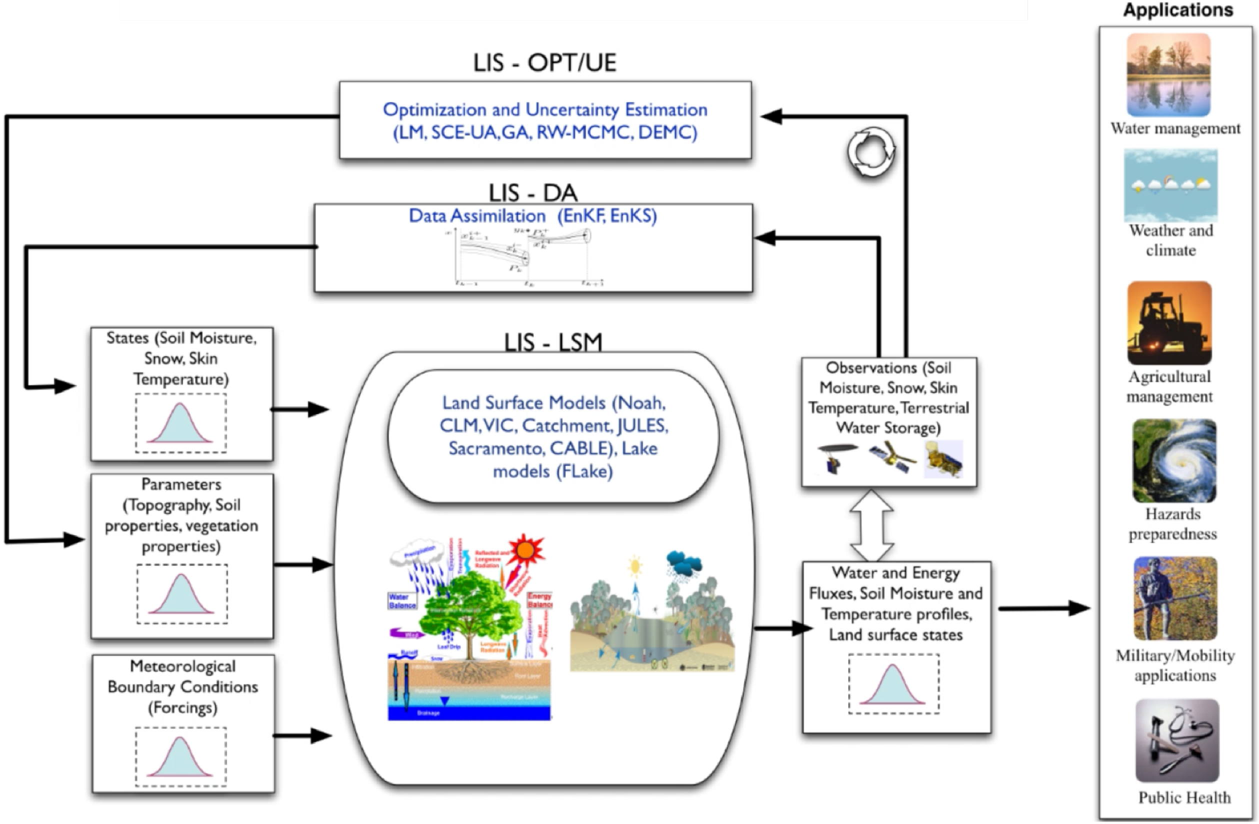
Implementation of Gaussian Processes in an Hydrological Model

[1,2]Jules Kouatchou, [1,2]Craig Pelissier, [3]Grey Nearing, [1]Dan Duffy
[1]Christa D Peters-Lidard and [1]Jim Geiger

[1]NASA Goddard Space Flight Center, Maryland, USA
[2]Science Systems and Applications Incorporated, Maryland, USA
[3]University of Alabama, Alabama, USA

LIS Model



Applications
LIS - OPT/UE
Optimization and Uncertainty Estimation
(LM, SCE-UA,GA, RW-MCMC, DEMC)
LIS - DA
Data Assimilation (EnKF, EnKS)
States (Soil Moisture, Snow, Skin Temperature)
Parameters (Topography, Soil properties, vegetation properties)
Meteorological Boundary Conditions (Forcings)
LIS - LSM
Land Surface Models (Noah, CLM,VIC, Catchment, JULES, Sacramento, CABLE), Lake models (FLake)
Observations (Soil Moisture, Snow, Skin Temperature, Terrestrial Water Storage)
Water and Energy Fluxes, Soil Moisture and Temperature profiles, Land surface states
Water management
Weather and climate
Agricultural management
Hazards preparedness
Military/Mobility applications
Public Health

Gaussian Regression Process

- Nonparametric and probabilistic model
- Flexible with the capability to adapt the model complexity
- Training data is not summarized by few parameters
- Probabilistic nature allows a structured way of capturing the uncertainties in both the model itself and the measured data.

$$p(\mathbf{f} \,|\, \mathbf{x}, \theta) = \mathcal{N}\left(\mathbf{0}, \,|\, K(\mathbf{x}, \mathbf{x}', \theta)\right)$$

$$K(\mathbf{x}, \mathbf{x}', \theta) = \sigma_f^2 \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}')\Sigma^{-1}(\mathbf{x} - \mathbf{x}')\right]$$

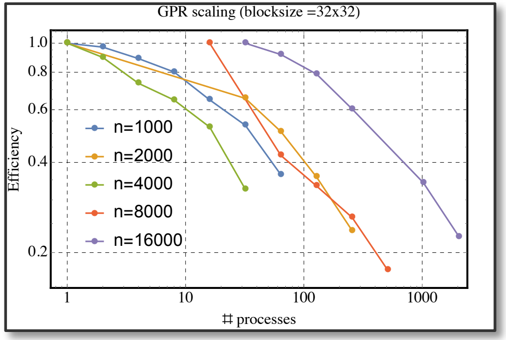
$$p(\mathbf{y} \,|\, X, \theta) = \int p(\mathbf{y} \,|\, \mathbf{f}, X, \theta) p(\mathbf{f} \,|\, X, \theta) d\mathbf{f}$$

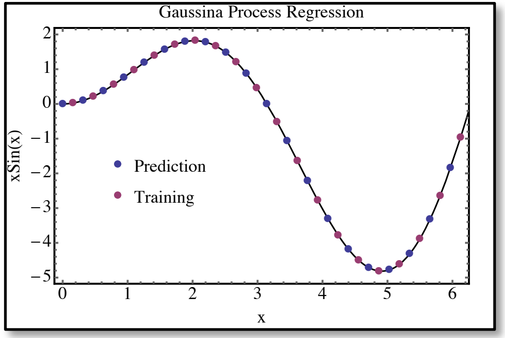
$$p(\mathbf{f}_* \,|\, \mathbf{y}, X, \theta) = \mathcal{N}\left(\bar{\mathbf{f}}_*, \mathrm{Cov}(\mathbf{f}_*)\right)$$

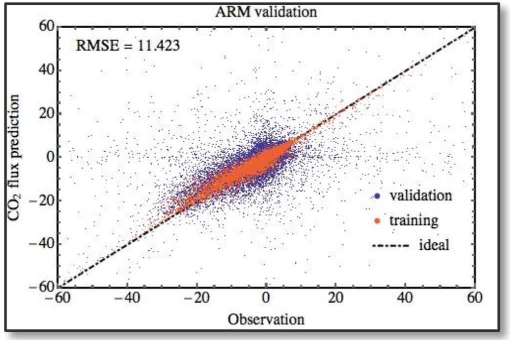
Computational Aspects of GPR

- **GPR can be expensive**
- **Prediction time can exceed models --- comparable with LIS models.**

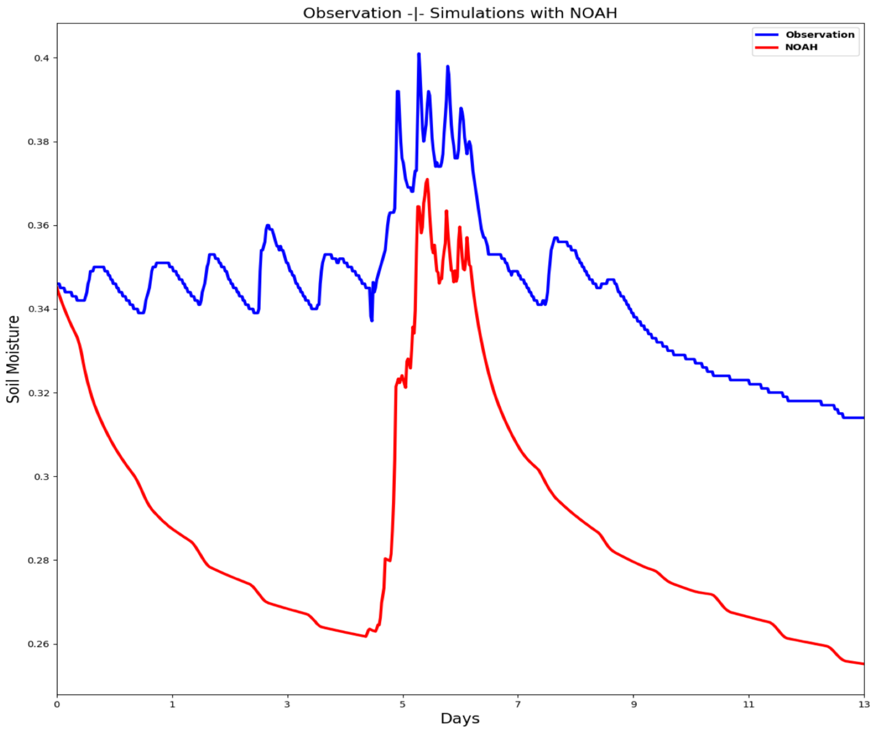
- **Training**
  - Dominated by DGEMM and Inversion.
  - Matrix size **N = #samples** --- limited to a few 10k.
  - Usually several 1000s of DGEMM/Inversions
- **Prediction**
  - 1 N X M Matrix-Vector multiplication. N = # samples M= # prediction points.
  - Dominated by matrix initialization (exp() function).
- **Implementation**
  - Parallelization --- ScaLapack/MKL
  - C++ and Fortran (interface only)







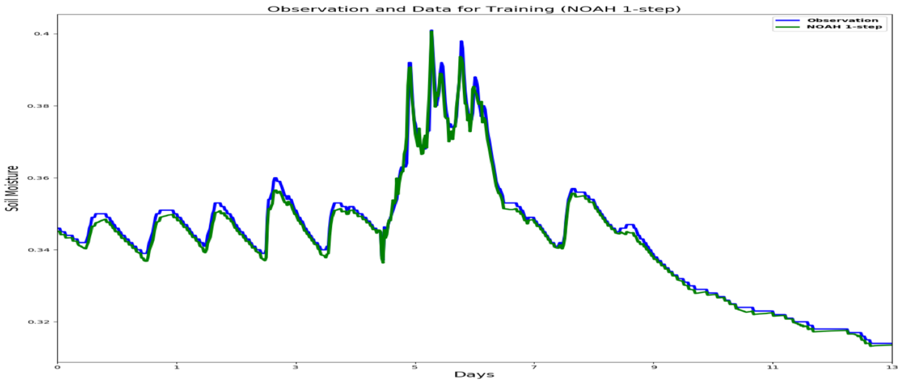
Machine Learning (GPR) in LIS

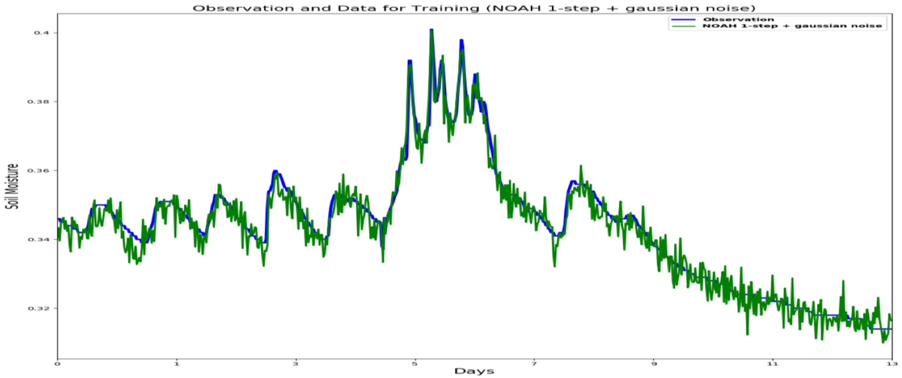


1. Develop a ML algorithm (Gaussian Process Regression) that trains on observational data and is HPC enabled.
2. Create in LIS a Land Surface Model option that calls the ML subroutines and returns the required output.
3. Train (offline) on field data with data assimilation or lagged forcing data.
4. Compare the prediction accuracy with other models such as NOAA.

Objective: ML model fills the gap between Observation and NOAH (see figure on the left)

Training Data

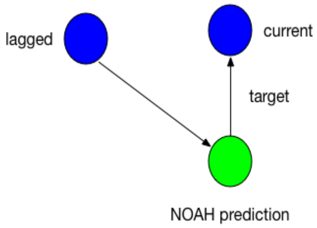


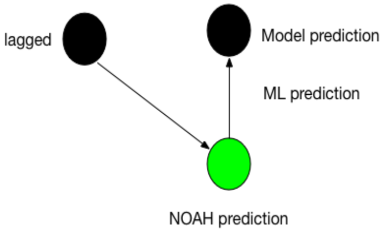


1. For a period of three years, ran one timestep at the time NOAH by starting with soil moisture from observation. Repeated several times till equilibration (NOAH 1-step).
2. This allows us to capture dynamics in NOAH
3. ML training is done on the NOAH deviation from observation

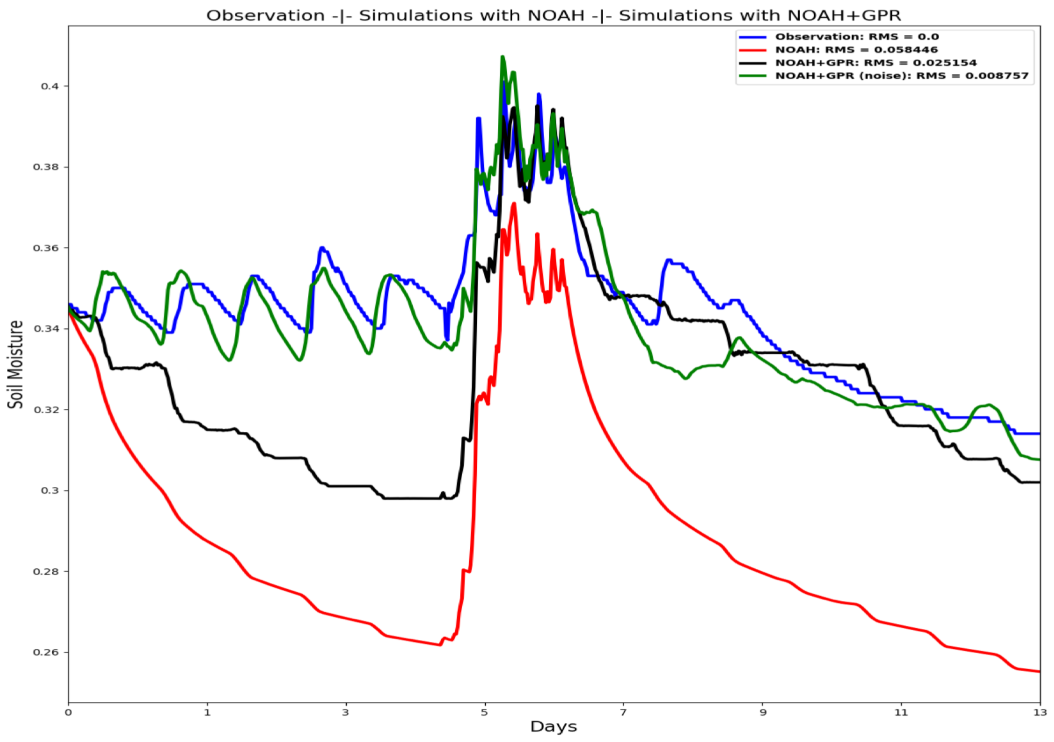
Two sets of training samples:

- **Set 1**: described above (see top figure on the left)
- **Set 2**: Complete the NOAH 1-step and add a Gaussian noise to dry data points only. (see bottom figure on the left)





Validation Results



1. Ran NOAH with soil moisture from observation as initial value
2. The NOAH predicted soil moisture (along with 7 other parameters) is fed into the GPR to produce the deviation from observation
3. The new NOAH soil moisture value is the sum of the predicted one and the deviation

Results

- The introduction of the GPR leads to better calculations of soil moisture (black plot).
- Adding Gaussian noise to samples significantly improves the prediction (green plot).